Demolingüística, Internet i dades massives

Natxo Sorolla[1], Àlex Nobajas[2] i Jordi Morales i Gras[3]
CUSC–UB[1], Universitat Rovira i Virgili[1], Xarxa CRUSCAT-IEC[1], Keele Univsersity[2], Euskal Herriko Unibertsitatea[3]
natxosorolla@urv.cat

Recepció: 26/04/2017, acceptació: 21/08/2017

Resum: L'article revisa les novetats que aporten a la demolingüística les noves dades massives disponibles amb la proliferació d'Internet, i sobretot de les plataformes digitals de xarxes socials. Explora els avantatges i inconvenients d'aquesta nova font de dades. Es revisen treballs desenvolupats sobre l'ús de les llengües en SMS, correus electrònics, xats, fòrums, blocs, Wikipedia o Facebook, i es posa especial atenció en Twitter. L'anàlisi sociolingüística de les dades massives es dirigeix a les tries lingüístiques, un objectiu propi de la demolingüística, però també a l'anàlisi més estrictament lingüística, com l'anàlisi de sentiment, el variacionisme o la dialectologia. Entre les possibilitats d'expansió de la demolingüística en dades massives es destaquen les possibilitats de l'anàlisi de les interaccions socials.

Mots clau: demolingüística, Internet, dades massives, xarxes socials, Twitter, sociolingüística catalana

Demolingüística, Internet y macro datos

Resumen: El artículo revisa las novedades que aportan a la sociolingüística los nuevos macro datos disponibles con la proliferación de Internet y, sobretodo, de las plataformas digitales de redes sociales. Explora las ventajas y los inconvenientes de esta nueva fuente de datos. Se revisan los trabajos desarrollados sobre el uso de las lenguas en SMS, correos electrónicos, chats, fórums, blogs, Wikipedia i Facebook, y se pone especial atención en Twitter. El análisis sociolingüístico de los macro datos se dirige hacia las elecciones lingüísticas, como el análisis de sentimiento, variacionismo o la dialectología. Entre las posibilidades de expansión de la demolingüística en macro datos se destacan las posibilidades de análisis de las interacciones sociales.

Palabras clave: demolingüística, Internet, macro datos, redes sociales, Twitter, sociolingüística catalana.

Demolinguistics, internet and big data

Abstract: The article reviews the how big data and digital social networks can contribute to demolinguistic research and it explores the pros and cons these new data sources offer. Existing research on the use of languages in SMS, emails, chats, forums, blogs, Wikipedia or Facebook is explored, while the Twitter case is studied in more detail. The sociolinguistic analysis of big data is focused on language choices, an intrinsic goal of demolinguistics, but also to linguistic analysis, such as sentiment analysis, linguistic variation or dialectology. Among the different possibilities using big data for demolinguistic research offers, we highlight the characteristics of analyzing social interactions.

Key words: demolinguistics, Internet, big data, social networks, Twitter, Catalan sociolinguistics.

1. Introducció

La manera més popular d'estudiar els aspectes demolingüístics d'una societat han estat les enquestes i censos, i de manera molt més residual, les observacions massives. Aquests mètodes continuaran sent molt útils, perquè donen una informació molt rica, que s'ajusta a les necessitats de les recerques que s'estan duent a terme. Però pateixen d'una sèrie de limitacions importants. En primer lloc, els estudis demolingüístics més popularitzats es basen sobretot en dades declarades pels mateixos entrevistats.[1] En segon lloc, fer recerca amb aquests mètodes costa molts diners, ja que cal involucrar una quantitat important de gent perquè les mostres siguin representatives de la societat que es vol estudiar. En tercer lloc, cal bastant temps per dur-les a terme, per tant és molt difícil de tenir informació de manera continuada. Aquest fet, lligat a l'alt cost d'aquestes tècniques, implica que la seua periodicitat no pot ser gaire alta. Aquestes limitacions han implicat que tradicionalment només les institucions públiques o amb recursos han pogut recollir dades demolingüístiques de manera periòdica i general, i els investigadors hi han pogut accedir amb major o menor facilitat segons les polítiques de transparència i accessibilitat als estudis públics.

2. Demolingüística de les dades massives: avantatges i inconvenients

El panorama demolingüístic, però, està canviant. Els canvis tecnològics que s'han produït des de la dècada dels 70 han conformat una revolució en la generació del coneixement, el processament de la informació i la gestió de la comunicació, que està tenint conseqüències socials similars a les de la Revolució Industrial (Castells 1996). Les noves tecnologies de la comunicació i les noves maneres amb les quals la gent es relaciona digitalment permeten a les ciències socials i humanes tenir accés per primer cop de manera massiva a textos, interaccions i converses espontànies. La popularització de les xarxes socials virtuals, i la consegüent aparició de les dades massives —conegudes popularment pel seu nom en anglès, *big data*— ha permès que les dades sociolingüístiques es puguin recollir també emprant aquestes noves tècniques. Mitjançant l'obtenció massiva d'interaccions en línia es tenen a l'abast milions de converses espontànies entre persones, que d'altra manera seria impossible d'obtenir. A més a més, aquesta nova font de dades permet evitar els problemes abans esmentats, ja que un cop el sistema està programat, el cost per cada nova captura de dades és negligible, i permet obtenir dades de manera continuada i de qualsevol àmbit geogràfic, evitant així els problemes de continuïtat i de cobertura.

Tot i això, no totes les xarxes socials són adients per obtenir dades d'aquesta manera. Per exemple, la que actualment té una presència més universalitzada, Facebook,[2] no ofereix gaires dades en obert a causa de les seues polítiques de privadesa, mentre que d'altres com Instagram o Snapchat són eminentment visuals, i el component lingüístic és menys prominent. Altres espais comunicatius virtuals, sovint anomenats *2.0* —pel fet de demanar participació d'aquell usuari, el qual en les anomenades *1.0* es denominava *navegant*—, com Tripadvisor, blogs o tota mena de fòrums, sí que es poden utilitzar com a fonts de dades

[1] Vegeu les reflexions de Fabà i Rosselló, i Iurrebaso, en els articles d'aquest mateix monogràfic sobre demolingüística de *Llengua, Societat i Comunicació* <http://revistes.ub.edu/index.php/LSC>.

[2] A Catalunya, per exemple, el 57,7% de la població ha utilitzat alguna xarxa social virtual en els últims tres mesos, dels quals la primera posició l'ocupa Facebook, amb el 79,2% d'usuaris, i la segona Twitter, amb un 9,6% (CEO, 2015).

sociolingüístiques, sovint geolocalitzades, però la seua especialització temàtica els allunya dels plantejaments demoscòpics generals. És per això que bona part dels investigadors fan servir Twitter com a font de dades, ja que ofereix les dades de manera relativament senzilla i es basa principalment en informació de caràcter textual. Twitter permet escriure als usuaris comentaris curts de fins a 140 caràcters, anomenats *piulades* o *tuits*, que qualsevol usuari d'Internet generalment pot veure, i qualsevol usuari de Twitter pot comentar. Això dona lloc al fet que tot i la seua explícita limitació en la longitud dels missatges, es puguin establir converses i interaccions entre els usuaris. Algunes institucions o polítics del país apleguen un bon nombre de seguidors, com els 200.000 seguidors de Carles Puigdemont o els 105.000 de la Generalitat Valenciana.

Actualment Twitter ofereix, per descarregar de manera gratuïta, un 1% aproximadament de totes les piulades que es generen, una xifra que pot semblar minsa però tenint en compte el volum total d'interaccions representa una quantitat ingent d'informació. Hi ha opcions per obtenir més dades, com per exemple la coneguda com a *decahose*, que proporciona el 10% de les piulades, però són costoses i per a estudis de caràcter sociolingüístic no solen ser necessàries. Amb tot, quan el criteri és temàtic, acostuma a ser assumible econòmicament la recopilació mitjançant eines privatives, tals com capturar la totalitat de piulades al voltant d'una sèrie de paraules clau (com per exemple *tornada*, *escola* i *Lleida*) o etiquetes (*hashtags*, en anglès).

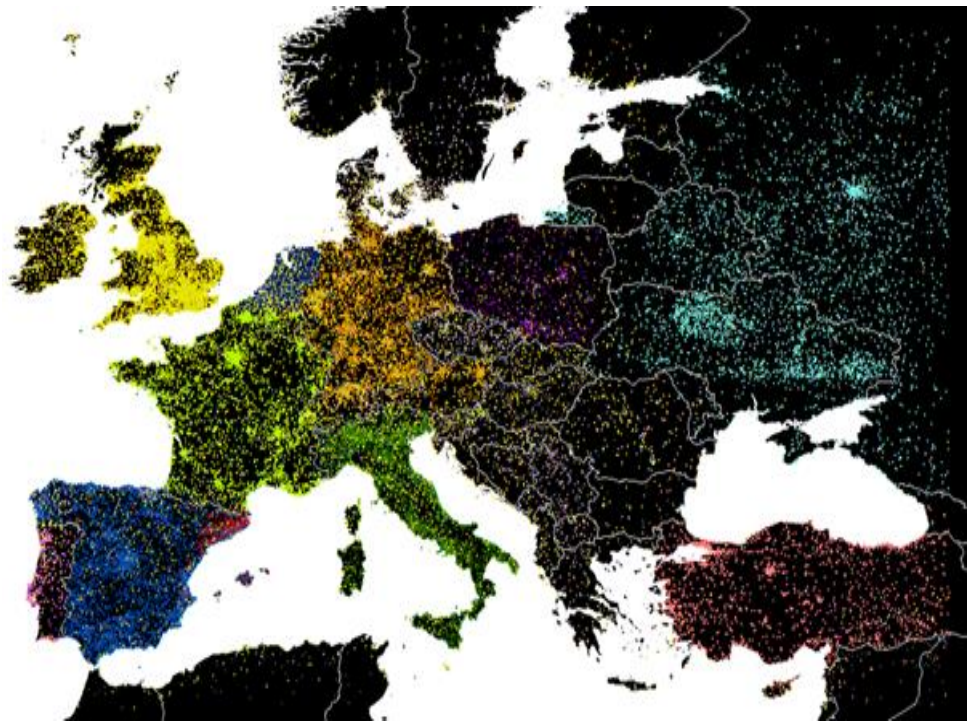
Encara que siguin dades fàcils d'aconseguir, cal tenir en consideració una sèrie de limitacions abans de fer servir dades provinents de Twitter. En primer lloc, cal tenir en compte que la major part dels usuaris actualment es concentra en població adulta-jove, és més comú entre població amb nivells formatius superiors, i també entre homes. Això és un fet que el distingeix d'altres xarxes, com Facebook, que és la xarxa majoritària en tots els grups d'edat, i per tant està menys concentrada generacionalment, o Instagram, concentrada en població més jove, o Linkedin, xarxa més significativa en població en edat laboral.[3] A més, cal tenir en compte dues qüestions cabdals intrínseques a Twitter: es tracta d'usos escrits, i habitualment sense restriccions d'accés per a qualsevol persona. Aquests dos fets en contextos de minorització lingüística poden fer modular les tries lingüístiques, tant pel que fa a les competències escrites de la població com al públic potencial difús, sobretot en casos de minorització. Per tant, cal tenir en compte que els estudis sobre Twitter analitzen comportaments lingüístics escrits dirigits a un públic difús i de sectors poblacionals joves amb nivells formatius superiors i masculinitzats. Per una altra banda, no totes les zones del món tenen la mateixa quantitat d'usuaris a Twitter, ja que hi ha fins i tot indrets on el seu ús està prohibit. En el context europeu, però, la major part dels països tenen una proporció important de la seua població que és usuària de Twitter. També cal tenir en compte que la limitació a 140 caràcters per piulada de vegades no permet una detecció idiomàtica precisa, a més del fet que no totes les piulades inclouen informació geogràfica i que en alguns casos els usuaris poden usar l'espai sobre la localització per mostrar altres identitats o informacions diferents.

Tot i això, aquestes noves tecnologies ajuden a preveure un futur engrescador per a la recerca sociolingüística. La possibilitat d'aconseguir milions de converses espontànies, sense la potencial interferència que pot ocasionar el fet de saber que s'està entrevistant o registrant, suposa un canvi de paradigma. Si a això li afegim que els mitjans actuals ens permeten detectar l'idioma en què estan fets els tuits de manera automàtica o les possibilitats que ofereixen les noves tècniques per fer mineria de dades, fins i tot sobre l'estructura de la llengua, resulta en el fet que podem analitzar quantitats ingents de dades. Per acabar-ho d'adobar, moltes de les piulades incorporen algun tipus d'informació sobre la localització de l'usuari que ha fet la piulada o des d'on l'ha fet. Això permet crear mapes

[3] Elaboració pròpia a partir de CEO 2015.

lingüístics de manera gairebé instantània (Figura 1), cosa que obre un món de possibilitats per a subdisciplines de la sociolingüística, com la geolingüística i la dialectologia.

FIGURA 1. Mapa lingüístic d'Europa creat a partir de dades provinents de Twitter. Cada punt representa una piulada i cada color l'idioma en què s'ha fet



Font: Elaboració propia (autor: Àlex Nobajas)

El tipus de tècniques d'anàlisi que es poden aplicar a dades massives virtuals són diverses i riques, i cobreixen plantejaments descriptius, exploratoris i inferencials. La telefonia mòbil ja despertà interès (socio)lingüístic amb els canvis que es produïen en l'estructura de la llengua escrita amb la difusió dels missatges de text de mòbil, els SMS (Anis 2007). L'interès continuà amb l'Internet més tradicional, amb recerques sobre la llengua de les webs (Pimienta et al. 2009, W3Techs 2017) i les seues relacions mitjançant els enllaços que mantenen (Ford i Batson 2011), dibuixant-se unes constel·lacions de grans llengües, envoltades per d'altres de mitjanes i menudes, a mode de la proposta teòrica de Swaan sobre el sistema mundial de llengües (2001). També, des d'aquesta perspectiva, es poden incloure les recerques sobre el consum de premsa digital (Negredo et al. 2016). Però va ser sobretot amb l'emergència d'Internet com a eina de comunicació que es despertà el màxim interès sociolingüístic sobre les NTIC, sovint des d'una perspectiva interaccional, ja fora amb les llistes de correu electrònic (Durham 2003), els xats (Paolillo 2002, Bergs 2006), les tries lingüístiques en fòrums de discussió (Androutsopoulos 2007) o els espais de realitat virtual (Axelsson et al. 2007).

La difusió del que es denomina *Internet 2.0*, on l'antic navegant es convertia en usuari i productor de continguts, va fer esclatar el nombre de recerques interessades en el comportament sociolingüístic de l'internauta, com per exemple les pràctiques lingüístiques als blocs (Nilsson 2003), però sobretot amb les anàlisis que destacaven la diversitat lingüística de plataformes multilingües com la Wikipedia (Ortega 2009). De fet, la Viquipèdia constituí un cas d'estudi important, amb recerques que analitzaven la relació entre els factors socials i el volum d'articles produïts en les diferents edicions lingüístiques, on el català se situava entre les primeres vint llengües, un volum equilibrat amb el nombre de parlants d'altres llengües europees (Pellejero, Sorolla, Nogué 2011), l'observació de com es replicava una mateixa estructura relacional en els articles de diferents edicions

lingüístiques de la Wikipedia (Zlatic et al. 2006) o els comportaments diferenciats dels usuaris de cada edició lingüística, també la catalana (Information geographies at the Oxford Internet Institute 2013).

Però el màxim interès s'ha despertat amb la implantació de les plataformes per al manteniment de xarxes socials en línia. Facebook ha estat hegemònic l'última dècada com a xarxa social en línia, però l'alta protecció que els usuaris mantenen sobre els continguts que comparteixen fa difícil la seua anàlisi. Amb tot, sociòlegs com Andreas Wimmer han estudiat en un sentit més ètnic les interaccions entre universitaris nord-americans, a partir de l'etiquetatge en les fotografies que compartien a la plataforma, podent testar la tendència dels alumnes de Nord-amèrica a relacionar-se amb aquells amb qui compartien trets ètnics, i comprovant que aquesta tendència era dèbil quan es controlava per altres factors socials, com les identitats més locals, l'origen regional, l'estatus socioeconòmic, els gustos musicals compartits, cursar les mateixes disciplines, residir en llocs comuns o altres efectes relacionals com la reciprocitat i la transitivitat en les relacions.

Les característiques pròpies de Twitter, com a plataforma netament lingüística, de caràcter més públic, i amb la participació d'un gruix important d'usuaris, han convertit aquesta xarxa social en línia en l'estudi de cas més explorat acadèmicament. Destaquen visualitzacions com la de Fischer (2011), amb la mapificació de la presència mundial de les llengües en les piulades, i on es destacava la presència majoritària de l'anglès a Twitter (41,6%), seguida a distància pel castellà (11,2%), el portuguès (9,5%) o l'indonesi (7,3%) (Leetaru et al. 2013). La major part dels països observaven una forta homogeneïtat lingüística interna, normalment amb una única llengua hegemònica, i algunes petites zones isolades amb una altra llengua. Però es destacava la diversitat interna a l'Europa central, els Balcans, el Líban, Israel i Cisjordània. Altres estudis generals com el de Mocanu et al. (2012) sobre els usos lingüístics a Twitter, destacaven la posició número 19 del català, entre les més usades a Twitter, i aprofundien en les peculiaritats geogràfiques de l'ús del català i el castellà a Catalunya, i la seua comparació amb estudis demolingüístics tradicionals.

Altres estudis com el de Takhteyev et al. (2012) es fixaven en els factors socials que afecten la formació de llaços entre usuaris de Twitter, tals com la llengua, la distància geogràfica, les fronteres administratives o les interconnexions aèries entre àrees geogràfiques, destacant que els efectes de compartir llengua són controlats per altres efectes més rellevants, com compartir país, i que el paper de l'anglès com a llengua franca matisa els efectes de les àrees lingüístiques sobre la formació de relacions entre usuaris. Saito i Masuda (2014) destacaven estructures diferents de popularitat a Twitter, que es reproduïen de manera diferent en cada llengua. En el cas de les llengües minoritzades, estudis com el de Jongbloed-Faber et al. (2016) aprofundien en l'ús digital del frisó, un fet molt rellevant per a les comunitats minoritzades, pel fet que el seu impacte digital es pren en consideració en la configuració de les polítiques lingüístiques de les empreses tecnològiques (Racó Català 2008).

Alguns sociolingüistes consolidats, com Li Wei, han participat en incursions aprofundides sobre els usos lingüístics a Twitter (Kim et al. 2014), analitzant els usos lingüístics a Twitter en tres contextos territorials multilingües (Suïssa, Quebec i Qatar) i analitzant la diversitat de tries lingüístiques dels usuaris, la màxima influència (nombre de seguidors) dels usuaris que trien les llengües locals (fins i tot en contextos on els usuaris d'anglès són majoritaris), i la màxima interacció dels usuaris monolingües entre ells. De fet, els investigadors destacaven que els usuaris que utilitzen llengües locals i anglès conformen un pontentre les diferents comunitats lingüístiques, i en conseqüència, l'anglès té un paper central com a llengua franca, fins i tot amb els usuaris monolingües d'anglès. A més, l'ús de l'anglès dels usuaris bilingües (llengua local i anglès) s'associava amb la densitat de seguidors que tenien en una o altra llengua, i els investigadors destacaven que es detectava una distribució funcional de la llengua segons les temàtiques tractades pels

usuaris bilingües, amb preeminència de les llengües locals en els temes informatius, polítics i debats dirigits a usuaris locals, i presència de l'anglès en temes sobre esdeveniments, viatges i oci, dirigits a usuaris exteriors. Altres investigadors (Ronen et al. 2014) han proposat que la influència mundial de les llengües no es basa en el volum de parlants o en el producte interior dels països on es parlen, sinó que és la posició de cadascuna en la xarxa que entrelliga les llengües. A partir de les traduccions de llibres, i de dades massives, com edicions a la Wikipedia en diferents llengües o missatges a Twitter que usen llengües simultàniament, es pot destacar l'anglès com a llengua franca global, però una distinció rellevant entre l'alemany, el francès i el castellà, que ocupen una posició de més interconnexió global de llengües, que no pas el xinès, l'àrab i l'hindi, que a pesar de tenir volums molt importants de parlants, ocupen un espai més perifèric en la interconnexió global de llengües. Segons l'estudi, aquesta posició d'interconnexió en la xarxa global de llengües és força rellevant a l'hora d'explicar la influència global d'una llengua, que es mesura a partir de les persones rellevants que tenen biografia en diverses llengües a la Wikipedia o apareixen a publicacions de referència.

En el cas específic de la llengua catalana, han estat poques les recerques desenvolupades, encara. Tölke (2015) va fer un estudi sobre l'ús del català a Twitter a la Marina Alta.

Més enllà de les tradicionals classificacions i anàlisis de correlacions, o la pròpia anàlisi del discurs que es produeix a Twitter sobre el concepte de *big data* (Suárez-Gonzalo, Guerrero-Solé 2016), també destaquen eines d'anàlisi semàntica. L'anàlisi de sentiment o mineria d'opinió persegueix la identificació d'informació subjectiva mitjançant procediments de lingüística computacional. Aquestes tècniques intenten detectar que la paraula *freda* és negativa quan es parla d'una pizza, però positiva quan es parla d'una llimonada, i són especialment eficients quan es desenvolupen amb eines d'aprenentatge automàtic (p.e. *machine learning*) capaces de generar algoritmes en les seues llibreries de processament del llenguatge natural. Hi ha una bona colla d'estudis que dediquen els esforços a mesurar el sentit dels missatges, sobre si són positius o negatius, i sobre la mesura de la felicitat (Dodds et al. 2011, 2014),[4] i les seues interaccions, per entendre l'aparent paradoxa que els amics sempre són més feliços que un mateix (Bollen et al.2016). Alguns investigadors, de fet, es dediquen a mesurar l'estat d'ànim en relació amb qüestions públiques, com el referèndum del Brexit al Regne Unit o les eleccions nord-americanes, i la capacitat de predicció dels resultats a partir de l'anàlisi de dades massives provinents de Twitter, Youtube o comentaris a la premsa electrònica (sense-eu.info Lord 2016). Altres estudis, amb participació d'investigadors valencians, intenten aprofundir en qüestions pròpies de la pragmàtica, tals com la possibilitat de detectar els missatges irònics o sarcàstics (Reyes et al. 2013, Sulis et al. 2016), destacant que l'ús d'emoticones ajuda poc a detectar els missatges irònics, però altres factors són molt més rellevants, com les incongruències entre la part inicial i la part final, les incoherències semàntiques, canvis d'escenari emocional, ús d'adverbis que intensifiquen el significat i l'ús de la hipèrbole o la tipografia.

En el camp del variacionisme, amb precedents sobre l'estudi de la variació de lèxic segons l'edat i el gènere dels blocs (Argamon et al. 2007),destaquen estudis com els de Dong Nguyen (2017) i altres investigadors als Països Baixos (Nguyen et al. 2013), que aprofundien en les possibilitats de predicció de l'edat i el gènere dels usuaris de Twitter a partir dels continguts que publicaven, destacant entre les variables que usen la distinció segons l'ús de les preposicions *jo* i *tu*, la llargada de les paraules, la llargada dels tuits, l'ús de paraules com *escola* o *fill*, l'ús d'emoticones o l'ús de paraules que indiquen suport, cortesia o desig.

[4] Vegeu l'extensa bibliografia i aplicacions a http://hedonometer.org/papers.html.

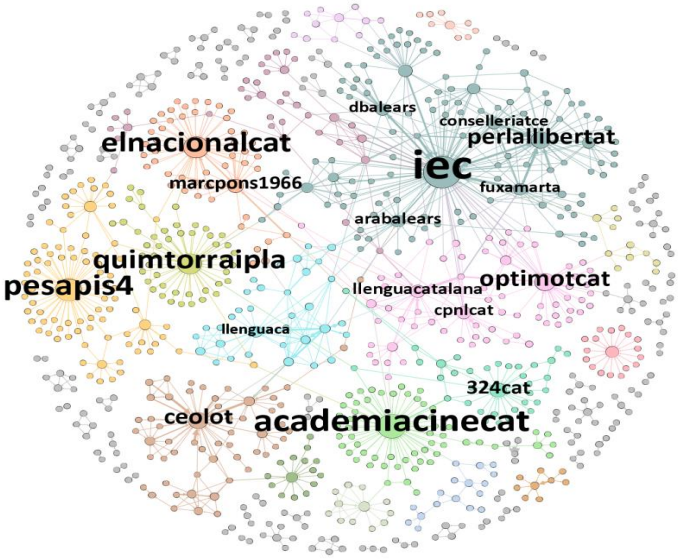
Pel que fa a la difusió de lèxic, més enllà de la detecció primerenca de desastres naturals a partir de la difusió de les etiquetes (elmundo.es. 2008, Amislove 2011), en el camp de la dialectologia Gonçalves i Sánchez (2014) utilitzaven els continguts de Twitter per estudiar el lèxic dialectal del castellà, apuntant a una gran varietat urbana que comparteixen les ciutats espanyoles i americanes, diferenciada d'una altra gran varietat rural, que es divideix en tres subvarietats: la d'Espanya, la d'Amèrica central i la d'Amèrica septentrional. En el cas de l'anglès, s'han utilitzat aplicacions de mòbil per dibuixar aquesta diversitat lingüística (Leeman 2016),[5] o altres investigadors l'han utilitzat per estudiar la difusió de lèxic dialectal (Russ 2012).

3. Perspectives de futur de les dades massives en demolingüística

Com hem vist, la recerca demolingüística basada en les dades massives pivota sobre dos eixos: l'anàlisi de les tries lingüístiques, que destaca els aspectes socials, i l'anàlisi de continguts, que se centra en els aspectes lingüístics, explotats des de camps diversos com el variacionisme, la dialectologia o la pragmàtica. Reprenent les notes que Fishman (1991) feia sobre la necessitat de recuperar els aspectes més sociològics de la recerca sociolingüística, destaquem la necessitat d'incorporar els aspectes interaccionals en la recerca demolingüística en dades massives, tant en el seu vessant més sociològic com en el lingüístic. L'anàlisi de xarxes socials és especialment pertinent per a la detecció d'espais de generació de discurs (Morales i Gras 2015, 2017) o de rols sociolingüístics i de processos comunicatius en general, i es pot aplicar sempre que tinguem dades sobre interaccions (p.e. A escriu a B en la llengua X, B escriu a C en la llengua Y, C escriu a A en la llengua X, etc.). Un exemple de l'aplicació de l'anàlisi de xarxes en sociolingüística, però que se situa encara en les dades de la demolingüística clàssica, és Sorolla (2016), en què es detecten diferents rols sociolingüístics dels individus a partir de les xarxes d'interacció, i que s'aplica a dades massives sobre interaccions, com les de Twitter. El vessant més lingüístic és el que mostra la Figura 2, on cada node és un usuari de Twitter que fa menció del concepte *llengua catalana*, cada aresta una menció dirigida d'un usuari a un altre, els usuaris més mencionats són més grans, i els colors dels nodes indiquen un grup d'usuaris que interactua amb major intensitat.

[5] Vegeu les recerques desenvolupades per David Britain i el seu equip: http://davebritain.weebly.com/publications.html.

Figura 2: Graf a partir de la monitorització del concepte llengua catalana a Twitter (entre el 25/1 i el 3/2 del 2017).



Font: Elaboració pròpia mitjançant Gephi (autor: Jordi Morales)

En la sociolingüística catalana encara no s'ha fet una adopció general de les metodologies per a la recerca de les dades massives. A banda d'alguns investigadors esparsos de Catalunya i el País Valencià que participen en projectes d'àmbit internacional, o mencions especials que es fan en recerques generals sobre les especificitats de la llengua catalana en l'espai virtual, bona part dels investigadors estan dispersos, i aquest àmbit de recerca és secundari en la seua carrera investigadora.[6]

Finalment, un punt clau a tenir en compte en l'anàlisi de dades massives virtuals és de caire metodològic, i consisteix en el control que exerceix la persona investigadora sobre l'objecte d'estudi. En aquesta fase embrionària de l'explotació de dades massives són habituals recerques, també demolingüístiques, amb manca de depuració dels continguts. Per exemple, obviar la presència de robots (*bots*) automàtics que fan publicacions automàtiques (de l'hora, per exemple) o usuaris que publiquen la seua geolocalització a partir de la ciutat del seu esportista favorit poden desvirtuar fàcilment alguns resultats. A més, cal tenir en compte que els investigadors sempre tindran dificultats per identificar variables estructurals, com el gènere, l'edat, el nivell d'estudis o el poder adquisitiu dels usuaris.[7] Per tant, cal integrar aquestes mancances en l'apropament metodològic i tècnic vers aquest tipus de dada. Tanmateix, això no deixa de ser una oportunitat per a plantejaments teòrics de caire relacional, focalitzats en els processos d'emergència i solidificació dels fenòmens socials i lingüístics, i que obren la porta a l'estudi directe d'interaccions empíriques —allò que la gent fa— al marge d'autocategoritzacions —allò que la gent diu que fa— que tan sovint ens condueixen a males interpretacions i biaixos explicatius. Les necessitats de l'Estat modern i les oportunitats tècniques del registre de dades varen fer nàixer la demografia i la sociologia modernes, i amb aquestes, la

[6] Un exemple és el monogràfic sobre les llengües en les tecnologies de la informació i la comunicació (TIC) de la revista *Treballs de Sociolingüística Catalana* (Societat Catalana de Sociolingüística 2016) fet mitjançant crida d'articles oberta.

[7] Per exemple, no es podrà conèixer amb la precisió que permeten les recerques demoscòpiques el nivell educatiu, la classe social, o ni tan sols el gènere o l'edat, d'un conjunt massiu de perfils de Twitter, tret que canviïn dràsticament les polítiques de privadesa de plataformes com Twitter i la legislació de protecció de dades, o s'articulin estratègies complexes de creuament de dades i microsegmentació (Nguyen, Gravel, Trieschnigg i Meder 2013).

demolingüística. Podran satisfer les dades massives les necessitats de la modernitat avançada, i podrà la demolingüística treure profit de les seues oportunitats tècniques?

4. Referències bibliogràfiques

AAMISLOVE (2011). *Tweets mentioning «earthquake» immediately following Virginia earthquake on 08/23/2011* [en línia]. <https://www.youtube.com/watch?v=XJ1EQbmJ_LQ> [Consulta: 9 octubre 2017].

ANDROUTSOPOULOS, JANNIS (2007). «Language choice and code-switching in German-based diasporic web forums». *The multilingual Internet: Language, culture, and communication online*, 340-361.

ANIS, JACQUES (2007). Neography: «Unconventional spelling in French SMS text messages». *The multilingual Internet: Language, culture, and communication online*, 87-115.

ARGAMON, S.; KOPPEL, M.; PENNEBAKER, J. W.; SCHLER, J. (2007). «Mining the Blogosphere: Age, gender and the varieties of self-expression» [en línia]. *First Monday*, *12*(9). <https://doi.org/10.5210/fm.v12i9.2003> [Consulta: 9 octubre 2017].

AXELSSON, ANE-SOFIE; ABELIN, ÅSA; SCHROEDER, RALPH (2007). «Anyone speak Swedish? tolerance for language shifting in graphical multi-user virtual environmnets» [en línia]. DANET, BRENDA; HERRING, SUSAN (eds.). *The multilingual Internet: Language, culture and communication online*. <http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195304794.001.0001/acprof-9780195304794-chapter-16> [Consulta: 9 octubre 2017].

BERGS, ALEXANDER (2006). «Analyzing online communication from a social network point of view: questions, problems, perspectives» [en línia]. *Language@Internet*, *3*(3). Retrieved from <http://www.languageatInternet.org/articles/2006/371> [Consulta: 9 octubre 2017].

BOLLEN, JOHAN; GONÇALVES, BRUNO; VAN DE LEEMPUT, INGRID; RUAN, GUANGCHEN (2016). «The happiness paradox: your friends are happier than you» [en línia]. *arXiv:1602.02665 [physics]*. <http://arxiv.org/abs/1602.02665> [Consulta: 9 octubre 2017].

CASTELLS, MANUEL (1996). *La societat xarxa* (L'era de la informació: economia, societat i cultura. Vol. I). Editorial UOC.

CEO (2015). *Xarxes socials i política catalana*. 2015 - REO 809 [en línia]. Centre d'Estudis d'Opinió (CEO), Generalitat de Catalunya. <http://ceo.gencat.cat/ceop/AppJava/pages/estudis/cerquesRapides/cercaAnys/fitxa/fitxaEstudi.html?colId=5670&lastTitle=Xarxes+socials+i+pol%EDtica+catalana.+2015&any=2016&rang1=&rang2=> [Consulta: 9 octubre 2017].

SHERIDAN DODDS, PETER; CLARK, ERIC M.; DESU, SUMA; MORGAN R., FRANK; REAGAN, ANDREW J.; RYLAND WILLIAMS, JAKE; MITCHELL, LEWIS; DECKER HARRIS, KAMERON; KLOUMANN, ISABEL M.; BAGROW, JAMES P.; MEGERDOOMIAN, KARINE; MCMAHON, MATTHEW T.; TIVNAN, BRIAN F.; DANFORTH, CHRISTOPHER M. (2014). «Human language reveals a universal positivity bias» [en línia]. *arXiv:1406.3855 [physics]*. <http://arxiv.org/abs/1406.3855> [Consulta: 9 octubre 2017].

SHERIDAN DODDS, PETER; KAMERON DECKER, HARRIS; KLOUMANN, ISABEL M.; BLISS, CATHERINE A.; DANFORTH, CRISTOPHER M. (2011). «Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter»[en línia]. *PLoS ONE*, *6*(12), e26752. <http://arxiv.org/abs/1101.5120> [Consulta: 9 octubre 2017].

DURHAM, MERCEDES (2003). «Language Choice on a Swiss Mailing List» [en línia]. *Journal of Computer-Mediated Communication*, *9*(1), 0-0. <https://doi.org/10.1111/j.1083-6101.2003.tb00359.x> [Consulta: 9 octubre 2017].

EL MUNDO (2008). «China tembló primero en Twitter» [en línia]. 12 de maig. <http://www.elmundo.es/elmundo/2008/05/12/catalejo/1210622213.html> [Consulta: 9 octubre 2017].

FISCHER, ERIC (2011). *Language communities of Twitter* i *Language communities of Twitter (European detail)* [en línia].
<http://www.flickr.com/photos/walkingsf/6277163176/>,
<http://www.flickr.com/photos/walkingsf/6276642489/> [Consulta: 9 octubre 2017].

FISHMAN, JOSHUA A. (1991). «Putting the 'socio' back into the sociolinguistic enterprise» [en línia]. *International Journal of the Sociology of Language*, *92*(1), 127-138. <https://doi.org/10.1515/ijsl.1991.92.127> [Consulta: 9 octubre 2017].

FORD, DANIEL; BATSON, JOSH (2011). «Languages of the World (Wide Web)» [en línia]. *Google Research Web*. <https://research.googleblog.com/2011/07/languages-of-world-wide-web.html> [Consulta: 9 octubre 2017].

GONÇALVES, BRUNO; SÁNCHEZ, DAVID (2014). «Crowdsourcing dialect characterization through Twitter» [en línia]. *PloS one*, *9*(11), e112074.

Information geographies at the Oxford Internet Institute (2013). *Geographic intersections of languages in Wikipedia* [en línia]. <http://geography.oii.ox.ac.uk/?page=geographic-intersections-of-languages-in-wikipedia> [Consulta: 9 octubre 2017].

IURREBASO, IÑAKI (2017). «Uso observado y declarado de la lengua. Aportación metodológica basada en el caso vasco» [en línia]. *Llengua, Societat i Comunicació* (monogràfic sobre demolingüística). [En premsa] <http://revistes.ub.edu/index.php/LSC> [Consulta: 9 octubre 2017].

JONGBLOED-FABER, LYSBETH; VAN DE VELDE, HANS; VAN DER MEER, COR; KLINKENBERG, EDWIN (2016). «Language Use of Frisian Bilingual Teenagers on Social Media» [en línia]. *Treballs de Sociolingüística Catalana*, 26. <http://revistes.iec.cat/index.php/TSC/> [Consulta: 9 octubre 2017].

KIM, SUIN; WEBER, INGMAR; WEI, LI; OH, ALICE (2014). «Sociolinguistic Analysis of Twitter in Multilingual Societies» [en línia]. *Proceedings of the 25th ACM Conference on Hypertext and Social Media* (p. 243–248). New York, NY, USA: ACM. <http://ingmarweber.de/wp-content/uploads/2014/07/Sociolinguistic-Analysis-of-Twitter-in-Multilingual-Societies.pdf> [Consulta: 9 octubre 2017].

LEEMANN, ADRIAN; KOLLY, MARIE-JOSÉ, PURVES, ROSS; BRITAIN, DAVID; GLASER, ELVIRA (2016). «Crowdsourcing Language Change with Smartphone Applications» [en línia]. *PLOS ONE*, *11*(1), e0143060. <https://doi.org/10.1371/journal.pone.0143060> [Consulta: 9 octubre 2017].

LEETARU, KALEV; WANG, SHAOWEN; CAO, GUOFENG; PADMANABHAN, ANAND; SHOOK, ERIC (2013) «Mapping the Global Twitter Heartbeat: The Geography of Twitter» [en línia]. *First Monday* 18, 5. 22 d'abril. <http://firstmonday.org/ojs/index.php/fm/article/view/4366> [Consulta: 9 octubre 2017].

LORD, DEBBIE (2016). «Who predicted the Trump win? A computer and a professor, that's who» [en línia]. *AJC*. 9 de noviembre. <http://www.ajc.com/news/national/who-predicted-the-trump-win-computer-and-professor-that-who/oH01mzXpRzSIESAXRHsBPP/> [Consulta: 9 octubre 2017].

MORALES I GRAS, J. (2015). Desenredando las identidades soberanistas vasca y catalana: un Análisis de Redes Sociales de las etiquetas de Twitter #BasquesDecide y #Up4Freedom. *Papeles del CEIC. International Journal on Collective Identity Research*, *2015*, 128.

—— (2017). *Soberanías enredadas: una perspectiva reticular, constructural y agéntica hacia los relatos soberanistas vasco y catalán contemporáneos en Twitter*. Euskal Herriko Unibertsitatea, Soziologia 2 Saila. <https://addi.ehu.es/handle/10810/22686> [Consulta: 9 octubre 2017].

MOCANU, DELIA; BARONCHELLI, ANDREA; GONÇALVES, BRUNO; PERRA, NICOLA; VESPIGNANI, ALESSANDRO (2012). «The Twitter of Babel: Mapping World Languages through Microblogging Platforms» [en línia]. *arXiv:1212.5238*. 20 de desembre. <http://arxiv.org/abs/1212.5238> [Consulta: 9 octubre 2017].

NEGREDO, SAMUEL; VARA-MIGUEL, ALFONSO; AMOEDO, AVELINO (2016). «Digital News Report .es 2016 Cambios decisivos en el consumo de noticias digitales» [en línia]. Center for Internet Studies and Digital Life (Universidad de Navarra). 21 de juny. <https://drive.google.com/file/d/0B2eyawMqcpTyLVpGRoNLQzAtcmc/view?pref=2&pli=1&usp=embed_facebook> [Consulta: 9 octubre 2017].

—— (2016). *Digital News Report.es 2016 Cambios decisivos en el consumo de noticias digitales* [en línia]. Center for Internet Studies and Digital Life (Universidad de Navarra). <https://drive.google.com/file/d/0B2eyawMqcpTyLVpGRoNLQzAtcmc/view?pref=2&pli=1&usp=embed_facebook> [Consulta: 9 octubre 2017].

NGUYEN, DONG (2017). *Text as social and cultural data A computational perspective on variation in text* [en línia]. <http://dongnguyen.nl/thesis.html> [Consulta: 9 octubre 2017].

NGUYEN, DONG PHUONG; GRAVEL, RILANA; TRIESCHNIGG, RUDOLF BEREND; MEDER, THEO (2013). «"How old do you think I am?" A study of language and age in Twitter» [en línia]. <http://eprints.eemcs.utwente.nl/23604/> [Consulta: 9 octubre 2017].

NILSSON, STEPHANIE (2003). «The function of language to facilitate and maintain social networks in research weblogs». *D-Essay. Umea Universitet, Engelska lingvistik. Retrieved February*, *27*, 2004.

ORTEGA SOTO, JOSÉ FELIPE (2009). *Wikipedia: A quantitative analysis* [en línia]. Universidad Rey Juan Carlos I, Madrid. <https://eciencia.urjc.es/bitstream/handle/10115/11239/thesis-jfelipe.pdf> [Consulta: 9 octubre 2017].

PAOLILLO, JOHN C. (2002). «Language variation on Internet Relay Chat: A social network approach» [en línia]. *Journal of Sociolinguistics*, *5*(2), 180-213. <https://doi.org/10.1111/1467-9481.00147> [Consulta: 9 octubre 2017].

PELLEJERO, B.; SOROLLA, N.; NOGUÉ PICH, M. (2011). La dimensió de les llengües a la Wikipedia i la seua relació amb els elements socials. Digithum: revista digital d'humanitats, 0(13), 37-49. <http://www.raco.cat/index.php/Digit/article/view/245133> [Consulta: 9 octubre 2017].

PIMIENTA, DANIEL; PRADO, DANIEL; BLANCO, ÁLVARO (2009). *Twelve years of measuring linguistic diversity in the Internet: balanceand perspectives* [en línia]. UNESCO, Information Society Division, Communication and Information Sector. <http://unesdoc.unesco.org/images/0018/001870/187016e.pdf> [Consulta: 9 octubre 2017].

RACÓ CATALÀ (2008). «La importància de fer notar a Google que es té el navegador configurat en català» [en línia]. 10 de juny. <https://www.racocatala.cat/noticia/17383/importancia-fer-notar-google-te-navegador-configurat-catala> [Consulta: 9 octubre 2017].

RONEN, SHAHAR; GONÇALVES, BRUNO; HU, KEVIN Z.; VESPIGNANI, ALESSANDRO; PINKER, STEVEN; HIDALGO, CÉSAR A. (2014). «Links that speak: The global language network and its association with global fame» [en línia]. *Proceedings of the National Academy of Sciences*, *111*(52), E5616-E5622. <https://doi.org/10.1073/pnas.1410931111> [Consulta: 9 octubre 2017].

ROSSELLÓ I PERALTA, CARLES DE; FABÀ, ALBERT (2017). «Dades declarades i observades, pros i contres» [en línia]. *Llengua, Societat i Comunicació* (monogràfic sobre demolingüística). <http://revistes.ub.edu/index.php/LSC> [Consulta: 9 octubre 2017].

RUSS, BRICE (2012). «Examining large-scale regional variation through online geotagged corpora» [en línia]. *ADS Annual Meeting*. <http://briceruss.com/ADStalk.pdf> [Consulta: 9 octubre 2017].

SAITO, KODAI; MASUDA, NAOKI (2014). «Two types of well followed users in the followership networks of Twitter» [en línia]. *PLoS ONE*, *9*(1), e84265. <https://doi.org/10.1371/journal.pone.0084265> [Consulta: 9 octubre 2017].

SOCIETAT CATALANA DE SOCIOLINGÜÍSTICA (2016). *Treballs de Sociolingüística Catalana, 26. Les llengües en les tecnologies de la informació i la comunicació (TIC)* [en línia]. Institut d'Estudis Catalans. <http://revistes.iec.cat/index.php/TSC/issue/view/5457/showToc> [Consulta: 9 octubre 2017].

SOROLLA, N. (2016, gener 14). *Tria de llengües i rols sociolingüístics a la Franja des de la perspectiva de l'anàlisi de xarxes socials*. Universitat de Barcelona, Departament de sociologia i anàlisi de les organitzacions. <http://www.tdx.cat/handle/10803/373905> [Consulta: 9 octubre 2017].

Suárez-Gonzalo, S.; Guerrero-Solé, F. (2016). «La conversación sobre big data en Twitter. Una primera aproximación al análisis del discurso dominante». *Comunicació: revista de recerca i d'anàlisi*, vol. 33 (2), p. 113-131. <http://revistes.iec.cat/index.php/TC> [Consulta: 9 octubre 2017].

SULIS, E.; HERNÁNDEZ FARÍAS, D. I.; ROSSO, P.; PATTI, V.; RUFFO, G. (2016). «Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not». *Knowledge-Based Systems*, *108*, 132-143.

SWAAN, ABRAM DE (2001). *Words of the World: The Global Language System*. Wiley.

REYES, A.; ROSSO, P.; VEALE, T.; SPRINGER, Ó. (2013). «A multidimensional approach for detecting irony in Twitter». *Language Resources and Evaluation*, 239-268.

TAKHTEYEV, YURI; GRUZD, ANATOLIY; WELLMAN, BARRY (2012). «Geography of Twitter networks» [en línia]. *Social Networks*, *34*(1), 73-81. <https://doi.org/10.1016/j.socnet.2011.05.006> [Consulta: 9 octubre 2017].

TÖLKE, VANESSA (2015). «L'ús de les llengües minoritàries en les xarxes socials. El valencià en Twitter». *Zeitschrift für Katalanistik*, *28*, 95-115.

W3Techs (2017). *Usage of content languages for websites* [en línia]. 13 de marc. <https://w3techs.com/technologies> [Consulta: 9 octubre 2017].

WIMMER, ANDREAS; LEWIS, KEVIN (2013). «Network boundaries». *Ethnic Boundary Making: Institutions, Power, Networks*, 139-173). New York: Oxford University Press.

ZLATIC, V.; BOZICEVIC, M.; STEFANCIC, H.; DOMAZET, M. (2006). «Wikipedias: Collaborative web-based encyclopedias as complex networks» [en línia]. *Physical Review E*, *74*(1). <https://doi.org/10.1103/PhysRevE.74.016115> [Consulta: 9 octubre 2017].